Published online 31 July 2009 *Nucleic Acids Research, 2009, Vol. 37, No. 12* **1–7**
doi:10.1093/nar/gkn000

The RNA workbench: Best practices for RNA and high-throughput sequencing bioinformatics in Galaxy

Björn A. Grüning [1,2,*], Jörg Fallmann [3], Dilmurat Yusuf [4], Sebastian Will [5], Anika Erxleben [1], Florian Eggenhofer [1], Torsten Houwaart [1], Bérénice Batut [1], Pavankumar Videm,[1], Andrea Bagnacani [9], Markus Wolfien [9], Steffen C. Lott [12], Youri Hoogstrate [10], Wolfgang R. Hess [12], Olaf Wolkenhauer [9], Steve Hoffmann [3], Altuna Akalin [4], Uwe Ohler [4,11], Peter F. Stadler [3,5,6,7], Rolf Backofen [1,2,8,*]

[1] Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Koehler-Allee 106, D-79110 Freiburg, Germany [2] Center for Biological Systems Analysis (ZBSA), University of Freiburg, Habsburgerstr. 49, 79104 Freiburg, Germany [3] Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstr. 16-18, 04107 Leipzig, Germany [4] Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125, Berlin, Germany [5] Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, 1090 Vienna, Austria [6] Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, 04103 Leipzig, Germany [7] Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, United States [8] BIOSS Centre for Biological Signaling Studies, University of Freiburg, Schänzlestr. 18, 79104 Freiburg, Germany [9] Department of Systems Biology and Bioinformatics, University of Rostock, Ulmenstr. 69, 18051 Rostock, Germany [10] Department of Urology, Erasmus University Medical Center, Wytemaweg 80, 3015 CN Rotterdam, Netherlands [11] Departments of Biology and Computer Science, Humboldt University, Unter den Linden 6, 10099 Berlin [12] Genetics and Experimental Bioinformatics, Faculty of Biology, University of Freiburg, Schänzlestr. 1, 79104 Freiburg, Germany

Received January 1, 2009; Revised February 1, 2009; Accepted March 1, 2009

ABSTRACT

RNA-based regulation has become a major research topic in molecular biology. The analysis of epigenetic and expression data is therefore incomplete if RNA-based regulation is not taken into account. Thus, it is increasingly important but not yet standard to combine RNA-centric data and analysis tools with other types of experimental data such as RNA-seq or ChIP-seq. Here, we present the RNA workbench, a comprehensive set of analysis tools and consolidated workflows that enable the researcher to combine these two worlds. Based on the Galaxy framework the workbench guarantees simple access, easy extension, flexible adaption to personal and security needs, and sophisticated analyses that are independent of command-line knowledge. Currently, it includes more than 50 bioinformatics tools that are dedicated to different research areas of RNA biology including RNA structure analysis, RNA alignment, RNA annotation, RNA-protein interaction, ribosome profiling, RNA-seq analysis and RNA target prediction. The workbench is developed and maintained by experts in RNA bioinformatics and the Galaxy framework. Together with the growing community evolving around this workbench, we are committed to keep the workbench up-to-date for future standards and needs, providing researchers with a reliable and robust framework for RNA data analysis.

Availability: The RNA workbench is available at https://github.com/bgruening/galaxy-rna-workbench

INTRODUCTION

Since recent advances in high-throughput sequencing (HTS) emphasized the importance and versatile role of (non-coding) RNAs, there is high demand for integrated computational analyses investigating RNA-mediated regulation. Previously existing workbenches (*miARma-Seq* (1) *RAP* (2), UEA Small RNA Workbench (**?**)) were focused on providing tools for the analysis of RNA deep sequencing data and do not contain RNA centric tools.

We addressed these needs by developing the *RNA workbench*. Based on the *Galaxy* framework (3) it combines a comprehensive set of tools for the analysis of RNA structures, RNA alignments, RNA-RNA and RNA-protein interactions, RNA sequencing, ribosome profiling, genome annotation, and many more. So far, we integrated more than 50 RNA-related tools, including suites like the ViennaRNA package, covering this broad variety of use-cases (a complete list of tools can be found on GitHub). Every available tool works as a single building-block that can be connected with other tools to create

[*] To whom correspondence should be addressed. Tel: +49 761 2037460; Fax: +49 761 2037462; Email: gruening@informatik.uni-freiburg.de or backofen@informatik.uni-freiburg.de

computational pipelines. Datasets can be incorporated in a similar manner, facilitating an intersection of diverse data sources such as DNA methylation with RNA-seq experiments. Input and output datasets can be defined by the user, and can be as diverse as the adapted set of tools. Established data types for sequence and/or structure information are accepted as input. Output data types follow the same principle, can be converted to different formats, or ultimately used to draw plots and create figures. The workbench provides tools for visualizations of RNA structure datasets, such as dot-bracket strings, and RNA 2D- or 3D structures. The workbench also covers a broad range of RNA secondary structure prediction and analysis tools such as RNAfold (4) or LocARNA (5, 6).

GOALS OF THE RNA WORKBENCH

The main driving force behind the development of the *RNA workbench* is the goal to establish a central, redistributable workbench for scientists and programmers working with RNA-related data, and build a sustainable community around it. This platform is unique in combining available tools, workflows and training material, as well as providing easy access for experimentalists. Simultaneously it serves as a central hub for programmers, which can easily integrate and deploy their existing or novel tools and workflows. The RNA workbench is based on three pillars: 1) a comprehensive set of RNA-bioinformatics tools, 2) easy and stable dissemination via *Galaxy* and *Docker*, and 3) a set of predefined workflows and associated descriptions/training material. The latter is needed for two reasons: First, it facilitates the use of the RNA workbench for researchers with limited bioinformatics experience, and second, it allows to integrate the workbench in the daily lab work by combining RNA-related analysis tasks with workflows for RNA-seq analysis.

Building on the shoulders of giants

In order to achieve long-term sustainability, we provide the essentials of our work on *BioConda* (https://bioconda.github.io) and *BioContainers* (http://biocontainers.pro) for reproducible deployments of tools into *Galaxy*. Using easy-to-distribute packages for all tool dependencies also enables automatic continuous integration tests for all developed tools and the workbench. After a tool passes the tests and gets accepted it will be made available via an automatic deployment into the *Galaxy ToolShed* (https://toolshed.g2.bx.psu.edu) (7). From the ToolShed, *Galaxy* administrators can easily install desired tools and workflows.

Easily accessible and reproducible analysis platform

For the fast dissemination of the *RNA workbench*, as well as for an easy integration with other HTS analysis tasks, we implemented the *RNA workbench* within the *Galaxy* framework. A major advantage of relying on *Galaxy* as the core framework is that it is possible to leverage its scalability, which enables the *RNA workbench* to run on single CPU installations as well as on large multi-node high performance computing environments. Furthermore, *Galaxy* provides researchers with means to reproduce their own workflows analyses, enabling them to rerun entire pipelines, or publish and share them with others. The *RNA workbench* is containerized, *i.e.*, administrators can deploy it via *Docker*. That makes it possible to have all tool installation dependencies already resolved, while still keeping maintenance tasks to a minimum. The provided layer of virtualization also allows the handling of user-defined input data in a secure and compartmentalized way, a key requirement for researchers working on sensitive data (*e.g.* patient data in clinics). Running the containerized *RNA workbench* simply requires installing *Docker* and starting the *Galaxy RNA workbench* image. Furthermore, containerizing *Galaxy* enables a customized *Galaxy* instance with a selected subset of tools dedicated to specific data analysis tasks, while keeping deployment and installation simple.

RNA-BIOINFORMATICS TOOLS

In its current state, the *RNA workbench* includes more than 50 tools covering all aspects of RNA research. In a community effort, these tools will be kept up-to-date and adapted to future needs. New tools will also be integrated and new ways to visualize data provided to the user. A current overview of tools available in the RNA workbench can be found at http://bgruening.github.io/galaxy-rna-workbench/.

In the following, we will highlight a few of the integrated tools.

The ***ViennaRNA*** package (4) consists of a suite of tools centered around the prediction of secondary structures of RNAs based on the thermodynamic Turner energy model. Thus, it covers prediction of optimal and suboptimal structures from single sequences as well as alignments, prediction of ensemble base pair probabilities, accessibility of sequences, and RNA-RNA interaction prediction. Importantly, predictions can be flexibly controlled by hard and soft structure constraints; the latter enables the inclusion of structure probing data.

AREsite2 (8) is a resource for the investigation of AU, GU and U-rich elements (ARE, GRE, URE) in human and model organisms. It provides information on genomic location, genomic context, RNA secondary structure context and conservation of annotated motifs in the whole gene body including introns. It is integrated into the RNA workbench via its REST interface, which provides search results directly in *Galaxy* for further analysis.

LocARNA (5, 6) provides a comparative analysis of multiple (unaligned) RNAs by simultaneous folding and alignment, implementing a fast variant of the Sankoff algorithm. Beyond pairwise and multiple alignments, it computes reliabilities of alignment columns and provides very fast analysis by simultaneous folding and matching. Finally, *LocARNA* supports anchor and structure constraints, which improve its applicability in practice.

doRiNA (9) is a database of RNA interactions in post-transcriptional regulation. The combined action of RNA-binding proteins (RBPs) and microRNAs (miRNAs) is believed to form the backbone of post-transcriptional regulation. *doRiNA* is implemented as data source tool inside the RNA workbench. This means that the *Galaxy* user is redirected to the post-transcriptional interaction database and can make selections using the optimized doRiNA interface.

Once the selection is done, the data is streamed directly to *Galaxy* and can be freely analyzed with other tools.

The ***Infernal*** (10) tool suite can construct probabilistic models, also called covariance models (CM), that represent the sequence and structure of an RNA family from a multiple sequence alignment with consensus secondary structure. The covariance model can be used to find more members of this RNA family via homology search.

PARalyzer (11) generates a high resolution map of interaction sites between RNA-binding proteins and their targets. The algorithm utilizes the deep sequencing reads generated by the PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) protocol. The use of photoactivatable nucleotides in the PAR-CLIP protocol results in more efficient crosslinking between the RNA-binding protein and its target relative to other CLIP methods; in addition a nucleotide substitution occurs at the site of crosslinking, providing for single-nucleotide resolution binding information. *PARalyzer* utilizes this nucleotide substitution in a kernel density estimate classifier to generate the high resolution set of protein-RNA interaction sites.

FuMa (12) can generate an integration report on predicted fusion genes from most RNA-seq fusion gene detection software. It automatically orders the result based on the frequencies of the fusion genes such that frequently predicted fusion genes can be extracted.

WORKFLOWS

One of the core concepts of the *RNA workbench* is the definition of standard workflows as a minimal set of building blocks around which a researcher can compose and tailor specific pipelines. For example, a researcher wants to analyze the effects of an RNA-binding protein (RBP) in regard to expression levels in wild-type compared to knockout or knockdown of the RBP of interest. In this case, one needs to combine the detection of differentially expressed genes in the two conditions with the information of publicly available CLIP-data, as provided for example by the *doRiNA* (9) database, to differentiate between direct and indirect targets. Workflows for the analysis of differentially expressed genes are part of the RNA-workbench, as well as an interface to *doRiNA*, such that it becomes an easy task to design a new workflow combining these analysis steps.

In *Galaxy*, workflows are typically created in two different ways: (1) from an existing history storing all tools applied in a previous analysis together with all pertinent parameters, or (2) from scratch, using a graphical editor via drag-and-drop of tools from the tool panel into the workflow editor. Within workflows, tools can be freely combined to ensure a maximum of flexibility in their usage and connectivity between different analysis steps, *e.g.* RNA structure analysis tools and RNA-seq data analysis. Various format converters embedded in *Galaxy* allow combining diverse analysis outputs. Easy sharing of workflows with other *Galaxy* users guarantees highly reproducible and transparent research. In other words, the workflows ensure that all analysis steps, tools and parameters of an experiment are documented and visible to researchers, readers and reviewers. Workflows can also be submitted to the *Galaxy ToolShed* or *myexperiment.org* (13) for further distribution. The *RNA workbench* currently includes publicly available standard workflows for RNA data analysis, *e.g.* for RNA-seq. These workflows contain all required steps such as quality control, mapping, differential expression analysis, and visualization of results. Provided workflows can easily be extended or modified, *e.g.* to use other read mappers available in *Galaxy*.

In the following, we will describe two sample workflows, one closely related to the detection of ncRNAs, which is a common task in RNA-related research. The other workflow is related to the analysis of RNA-seq data and is often needed as a subworkflow for more complex analysis tasks. These workflows are well annotated and described in the RNA workbench and extended by interactive *Galaxy tours*.

Analysis of (unaligned) non-coding RNAs

An important task is to test for the existence of a functional structure in a non-coding RNA. However, the secondary structure of structured non-coding RNAs is not significantly more stable compared to random sequences (14). Thus, putative functional structures can only be detected using information about conservation. Our workflow for non-coding RNAs performs the typical analysis steps required to detect conserved secondary structures, given a set of unaligned RNA sequences. It computes a sequence and a structure-based alignment by *MAFFT* (15) and *LocARNA*, respectively, and analyzes them with *RNAcode* (16) and *RNAz* (17) with appropriate parameter settings. *RNAz* and *RNAcode* both work on a given alignment. *RNAz* tests whether a consensus secondary structure is significantly conserved, whereas *RNAcode* differentiates coding from non-coding RNAs. Together these tools provide information, whether the RNAs are related and conserve a common secondary structure. In addition, a covariance model is built from the *LocARNA* alignment and subsequently used to search the given sequence database for RNAs with similar sequence- and structure-conservation. This workflow resembles the core of *RNAlien* (18), which is based on the same tools and is integrated into the RNA workbench. Going beyond the presented workflow, *RNAlien* automatically gathers sequences via homology search starting from a single sequence and constructs RNA family models in an iterative process.

To give an another example, in the context of μORFs detection, RNA-seq analysis, the identification of non-coding RNAs with *RNAcode* and *RNAz* and the detection of transcription start sites can be used to determine new, short transcripts that are expressed and do not exhibit secondary structure conservation (*i.e.*, are likely not functional ncRNAs). Subsequent analysis of Ribo-seq data can then provide additional evidence for a new transcript that may code for a small protein. For all these tasks, partial workflows and required tools are already integrated in our RNA-workbench, which implies that it is easy to set up a new workflow for a more complex task.

RNA-seq analysis: trimming, mapping and read count

As said before, the analysis of RNA-centric data like CLIP-seq requires the combination with other type of data, and very often RNA-seq. For that reason, we provide a standard RNA-seq workflow that can easily be combined with other workflows.The RNA-seq workflow (as shown in Figure 1) takes a list of RNA-seq datasets as input and successively executes a series of analysis steps - adapter & quality trimming, mapping to a reference genome and read count per annotated gene. The input allows two conditions *e.g.* treatment versus control and it also accepts single-end and paired-end reads for each condition. At the trimming step, the workflow employs *Trim Galore!* (19, 20) to perform adapter trimming. Then, *TopHat2* (21) is used to map the trimmed reads against the reference sequences, which should be provided by a user. As last step, the workflow executes *HTSeq-count* (22) to generate read counts per annotated gene for each condition and for each sequencing type. A reference annotation in Gene Transfer Format (GTF), *e.g.* provided by Ensembl (23) is required at this step. The final read counts can be used for the downstream assessment of differential expression using tools like *DESeq2* (24). The current workflow can serve as a template that can be modified by the user according to different needs, for instance, replacement of tools or modification of the wrapping strategy.

The main advantage in comparison to existing solutions such as *miARma-Seq* (1) and *RAP* (2) is that our RNA-workbench combines the realm of RNA-centric analysis on sequence and structure level with modern high-throughput sequence analysis. In this regard we provide well established tools for RNA structure prediction, analysis and visualization together with read mappers and expression analysis tools for HTS analysis. An other advantage is its flexibility in terms of available tools, visualization and the adaption and extension of predefined pipelines. The extensive feature set provided by the *Galaxy* framework allows for this flexibility. All tools that are available on the *Galaxy-Toolshed* can be installed along with their automatically resolved dependencies with a single click in the *Galaxy* interface. Best practice pipelines for the analysis of RNA-seq data are provided with the *Docker* image and easily be modified, extended or combined with other analysis pipelines via *Galaxys* workflow editor GUI.

IMPLEMENTATION

The workbench is implemented as portable virtualized container based on *Galaxy*. The *Galaxy* framework allows for reproducible and transparent scientific research which makes it easy to access, deploy and scale - conceptualized as a web service. The foundation of the workbench container is a generic *Galaxy Docker* instance (http://bgruening.github.io/docker-galaxy-stable/). On-top of this, pre-configured *Galaxy* tools can be automatically installed from the *Galaxy ToolShed* using the *Galaxy* API *BioBlend* (25). In *Galaxy*, tool dependencies are automatically resolved via BioConda, which is the bioinformatics channel for the Conda package manager. BioConda facilitates software packaging and enables installation at a user level, keeping track of different versions of the same software in virtual environments. These features are in line with the scope of *Galaxy*; maintaining large numbers of dependencies in a reproducible way. Therefore, all available tools within the RNA workbench are also distributed as BioConda packages and BioContainers, which are persistent, frozen, containerized versions of Conda packages. The RNA workbench ships with a variety of tools, tours, documentation, workflows and data that have been added as additional layers on top of the generic *Docker* instance. During development, the software has been tested extensively in a continuous integration setup (CI) at different levels: *Galaxy* itself, tool integration in *Galaxy* (IUC, galaxytools channels), dependencies (BioConda) and at the workbench level. Together with a strict version management on all levels, this contributes to a high degree of error-control and reproducibility. The RNA workbench started in January 2015 - with constant development over two years, and extensive testing in local and public *Galaxy* instances, such as the Freiburg Galaxy instance, the MDC instance in Berlin and Erasmus MCs Galaxian. More than 500 users accessed the RNA tools during the last two years and the virtualized *Docker* instance was already downloaded more than 500 times. Moreover, due to an open and transparent development process, there is a growing community that contributes to our workbench, which guarantees the sustainability of the RNA workbench project and maintenance of the underlying *Docker/rkt* images.

USING THE RNA WORKBENCH

Installation: The RNA workbench can be installed under OSX and Windows using the graphical tool Kitematic (https://kitematic.com) , or with the following Linux command:

```
docker run -d -p 8080:80 bgruening/galaxy-rna-workbench
```

This installation is production-ready and can be configured to use external computer clusters or cloud environments. Due to the very modular system, it is also possible to install all or only a few tools of the *RNA workbench* on available *Galaxy* servers. Just get in contact with your local *Galaxy* administrator. When using the *RNA workbench Docker* image, the user has full administration rights, which enables customization independent of potential user restrictions.

Training

For self-empowering the user, documentation and training of the *RNA workbench* are important. We included an extensive set of documentation in traditional formats, *e.g.* tool descriptions and README files.

We also provide training sessions around HTS data analyses and RNA-seq data analysis. The training materials ranging from the introduction to *Galaxy*, to usage and maintenance of *Galaxy* and the *RNA workbench* are freely accessible for self-paced studies at the Galaxyproject Github repository (http://galaxyproject.github.io/training-material). This training material is constantly improved and extended in an international community effort, including ELIXIR and EMBL. For HTS data analyses we provide training as a specific introduction to the topic with self-explanatory presentation slides, a hands-on training documentation describing the analysis workflow, all necessary input files ready-to-use via *Zenodo*, a *Galaxy Interactive Tour*, and a tailor-made *Galaxy Docker* image for the corresponding data analysis.

To provide an even more intense training experience within the *RNA workbench*, we also included interactive

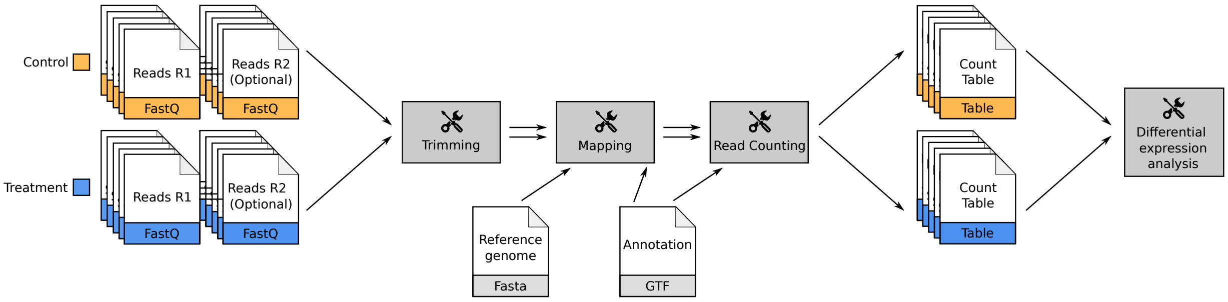

Figure 1. The workflow for analyzing RNA-seq data. The workflow tolerates single-end and paired-end reads derived from different conditions. It employs *TopHat2* for mapping and *htseq-count* to create the read counts. The final outputs contain read count per annotated gene for each condition and for each sequencing type.

training such as the *Galaxy Interactive Tours*. Such tours guide users through an entire analysis in an interactive and explorative way. It combines advantages from training videos and detailed protocols. Production of training videos is very time-consuming and tend to become outdated very soon, due to tool version changes or renewed workflows. In contrast to conventional screencasts, a *Galaxy Interactive Tour* can be easily updated and improved to guide the *Galaxy* user step-by-step *e.g.* through a whole HTS analysis starting from uploading the data to using complex analysis tools. Exemplary, the *RNA workbench* currently integrates two *Galaxy Interactive Tours*. The first one introduces a new user to the *Galaxy* interface and its usage with an RNA-seq example dataset. The second one illustrates secondary structure prediction of RNA molecules using parts of the *ViennaRNA* package. To show how *Galaxy Interactive Tours* can interactively guide users through the necessary steps of HTS analyses, the tours are also provided as online screencasts.

Visualization

Following data reduction as a key element of explorative research, there is a need for meaningful figures and visualizations that summarize results. The *RNA workbench* includes standard interactive plotting tools to draw bar charts and scatter plots from all kinds of tabular data and allows for connections to *Integrated Genome Browser* (27) and *UCSC* (28) like any other *Galaxy* instance. On top of this, we included three visualizations specific to RNA research. An interactive DotPlot visualization for secondary structures in EPS format (Figure 2b), a 2D visualization for the common dot-bracket format (Figure 2a) and a 3D visualization capable of visualizing PDB, SDF and MOL files containing three-dimensional coordinates (Figure 2c).

COMMUNITY

The *RNA workbench* project is an open source project that strives to create a community interested in accessible and reproducible RNA-related research. Knowing that real sustainability can only come true with a strong community we are aiming at more open participation, reward, and inclusion. We are working together with *Galaxy*, *BioConda*, *BioContainers* and *BioJS* and coordinating efforts to not reinvent the wheel but joining forces to create the new generation of bioinformatics infrastructure together. In the *RNA workbench* community, we practice the organizations on GitHub, IRC, and Gitter and welcome everyone to contribute on every level to improve the entire stack from documentation to tools and scientific workflows. Support will be provided through the same channels.

DISCUSSION

In this work we present the *RNA workbench*, maintained and developed by a constantly growing community. It provides a set of tools, each one being available as BioConda package as well as a Docker/rkt container (BioContainers). Based on the *Galaxy Docker* project, the proposed web server is more than the sum of its parts. It offers a comprehensive virtualized RNA workbench that can be deployed on every standard Linux, Windows and OSX computer, but can at the same time employ high-performance- or cloud-computing infrastructure. The presented workbench is unique in the sense that it combines RNA-centric analysis with other types of experiments. Major advantages of our approach to deliver a dockerized workbench for RNA centric analysis are the ease of installation, the high number of pre-included tools, the flexibility in regard to extension with other tools and workflows and the high reproducibility and transparency of workflows. The RNA workbench was designed as a community project, and as such it is easy for users to contribute to the workbench with workflows, new tools and training material, keeping the workbench up-to-date and valuable for research. Moreover, all components such as tools, workflows, visualizations, interactive tours and training material can be easily integrated into any available *Galaxy* instance for teaching, learning or exploratory purposes.

ACKNOWLEDGEMENTS

We thank the de.NBI and ELIXIR projects for supporting bioinformatics infrastructure. Thanks also to the *Galaxy* community, especially to the Freiburg *Galaxy* Team, for developing, maintaining and supporting this great framework. We also like to acknowledge the BioConda and BioContainers community for setting new standards in reproducible software deployments. Thanks also to the BioJS community for great discussions about scientific visualizations and how we can make them more accessible. Moreover, the authors acknowledge the support of many upstream developers that

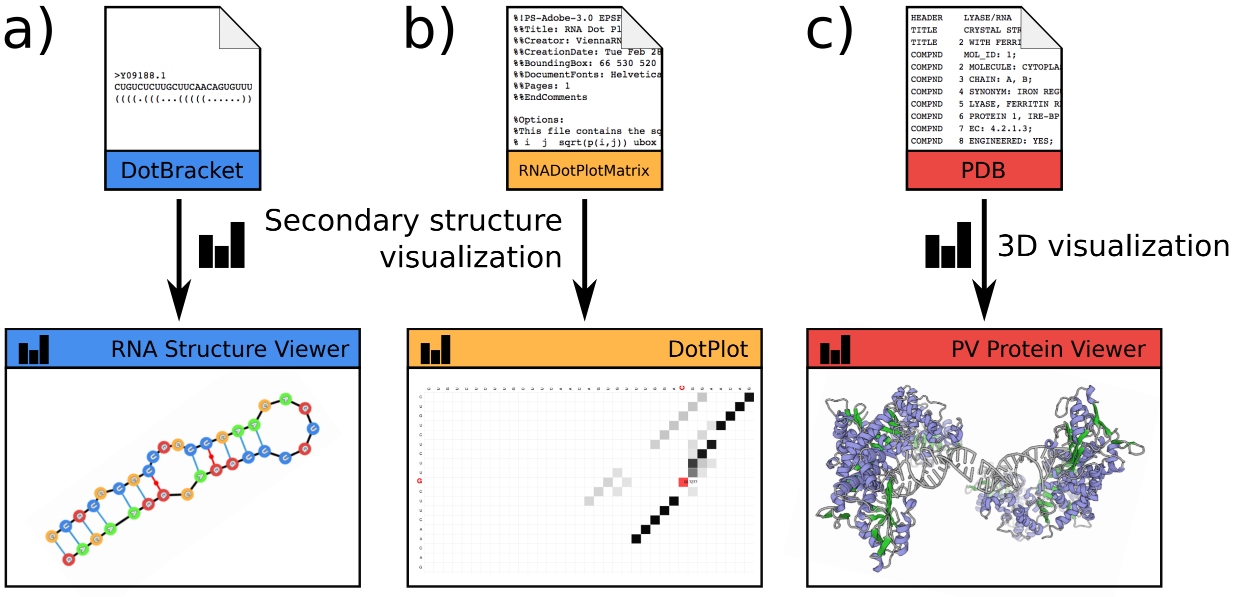

Figure 2. RNA structure visualization: The figure shows visualization for an *IRE1* RNA sequence, retrieved from Rfam database (26), via different backends integrated into the toolbox. (a) Secondary structure encoded in dot-bracket notation, can be displayed by the RNA structure viewer. (c) Base pairing probabilities are visualized as DotPlot. (c) Tertiary/Quaternary structure information encoded in protein-database format is rendered via Protein Viewer

helped us to integrate their tools into the *RNA workbench* and accepted patches.

Funding

The work is in part funded by Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and German Federal Ministry of Education and Research (BMBF grants 031 A538A/A538C RBC, 031L0101B/031L0101C de.NBI-epi, 031L0106 de.STAIR (de.NBI)). YH was funded within the framework of the Center for Translational Molecular Medicine (CTMM), TraIT project (grant 05T-401).

Conflict of interest statement. None declared.

REFERENCES

1. Eduardo Andrés-León, Rocío Núñez-Torres, and Ana M Rojas. miarma-seq: a comprehensive tool for mirna, mrna and circrna analysis. *Scientific reports*, 6, 2016.
2. Mattia D'Antonio, Paolo D'Onorio De Meo, Matteo Pallocca, Ernesto Picardi, Anna Maria D'Erchia, Raffaele A Calogero, Tiziana Castrignanò, and Graziano Pesole. Rap: Rna-seq analysis pipeline, a new cloud-based ngs web application. *BMC genomics*, 16(6):S3, 2015.
3. E. Afgan, D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. ?ech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, B. Gruning, A. Guerler, J. Hillman-Jackson, G. Von Kuster, E. Rasche, N. Soranzo, N. Turaga, J. Taylor, A. Nekrutenko, and J. Goecks. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.*, 44(W1):W3–W10, Jul 2016. [PubMed Central:PMC4987906] [DOI:10.1093/nar/gkw343] [PubMed:27137889].
4. R. Lorenz, S. H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. *Algorithms Mol Biol*, 6:26, Nov 2011. [PubMed Central:PMC3319429] [DOI:10.1186/1748-7188-6-26] [PubMed:22115189].
5. S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, 3(4):e65, Apr 2007. [PubMed Central:PMC1851984] [DOI:10.1371/journal.pcbi.0030065] [PubMed:17432929].
6. S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, and R. Backofen. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA*, 18(5):900–914, May 2012. [PubMed Central:PMC3334699] [DOI:10.1261/rna.029041.111] [PubMed:22450757].
7. D. Blankenberg, G. Von Kuster, E. Bouvier, D. Baker, E. Afgan, N. Stoler, J. Taylor, and A. Nekrutenko. Dissemination of scientific software with Galaxy ToolShed. *Genome Biol.*, 15(2):403, Feb 2014. [PubMed Central:PMC4038738] [DOI:10.1186/gb4161] [PubMed:25001293].
8. J. Fallmann, V. Sedlyarov, A. Tanzer, P. Kovarik, and I. L. Hofacker. AREsite2: an enhanced database for the comprehensive investigation of AU/GU/U-rich elements. *Nucleic Acids Res.*, 44(D1):D90–95, Jan 2016. [PubMed Central:PMC4702876] [DOI:10.1093/nar/gkv1238] [PubMed:26602692].
9. K. Blin, C. Dieterich, R. Wurmus, N. Rajewsky, M. Landthaler, and A. Akalin. DoRiNA 2.0–upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.*, 43(Database issue):D160–167, Jan 2015. [PubMed Central:PMC4383974] [DOI:10.1093/nar/gku1180] [PubMed:25416797].
10. E. P. Nawrocki and S. R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, Nov 2013. [PubMed Central:PMC3810854] [DOI:10.1093/bioinformatics/btt509] [PubMed:24008419].
11. D. L. Corcoran, S. Georgiev, N. Mukherjee, E. Gottwein, R. L. Skalsky, J. D. Keene, and U. Ohler. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.*, 12(8):R79, Aug 2011. [PubMed Central:PMC3302668] [DOI:10.1186/gb-2011-12-8-r79] [PubMed:21851591].
12. Y. Hoogstrate, R. Bottcher, S. Hiltemann, P. J. van der Spek, G. Jenster, and A. P. Stubbs. FuMa: reporting overlap in RNA-seq detected fusion genes. *Bioinformatics*, 32(8):1226–1228, Apr 2016. [DOI:10.1093/bioinformatics/btv721] [PubMed:26656567].
13. C. A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, and D. De Roure. myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res.*, 38(Web Server issue):W677–682, Jul 2010. [PubMed Central:PMC2896080] [DOI:10.1093/nar/gkq429] [PubMed:20501605].
14. E. Rivas and S. R. Eddy. Noncoding RNA gene detection using

comparative sequence analysis. *BMC Bioinformatics*, 2(1):8, 2001.

15. K. Katoh and D. M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30(4):772–780, Apr 2013. [PubMed Central:PMC3603318] [DOI:10.1093/molbev/mst010] [PubMed:23329690].
16. S. Washietl, S. Findeiss, S. A. Muller, S. Kalkhof, M. von Bergen, I. L. Hofacker, P. F. Stadler, and N. Goldman. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*, 17(4):578–594, Apr 2011. [PubMed Central:PMC3062170] [DOI:10.1261/rna.2536111] [PubMed:21357752].
17. A. R. Gruber, R. Neubock, I. L. Hofacker, and S. Washietl. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Res.*, 35(Web Server issue):W335–338, Jul 2007. [PubMed Central:PMC1933143] [DOI:10.1093/nar/gkm222] [PubMed:17452347].
18. F. Eggenhofer, I. L. Hofacker, and C. Honer Zu Siederdissen. RNAlien - Unsupervised RNA family model construction. *Nucleic Acids Res.*, 44(17):8433–8441, Sep 2016. [PubMed Central:PMC5041467] [DOI:10.1093/nar/gkw558] [PubMed:27330139].
19. F. Krueger. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries.
20. Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), 2011. [DOI:10.14806/ej.17.1.200].
21. D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14(4):R36, Apr 2013. [PubMed Central:PMC4053844] [DOI:10.1186/gb-2013-14-4-r36] [PubMed:23618408].
22. S. Anders, P. T. Pyl, and W. Huber. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, Jan 2015. [PubMed Central:PMC4287950] [DOI:10.1093/bioinformatics/btu638] [PubMed:25260700].
23. B. L. Aken, P. Achuthan, W. Akanni, M. R. Amode, F. Bernsdorff, J. Bhai, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, T. Juettemann, S. Keenan, M. R. Laird, I. Lavidas, T. Maurel, W. McLaren, B. Moore, D. N. Murphy, R. Nag, V. Newman, M. Nuhn, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, S. P. Wilder, A. Zadissa, M. Kostadima, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, F. Cunningham, A. Yates, D. R. Zerbino, and P. Flicek. Ensembl 2017. *Nucleic Acids Res.*, 45(D1):D635–D642, Jan 2017. [PubMed Central:PMC5210575] [DOI:10.1093/nar/gkw1104] [PubMed:27899575].
24. M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550, 2014. [PubMed Central:PMC4302049] [DOI:10.1186/s13059-014-0550-8] [PubMed:25516281].
25. C. Sloggett, N. Goonasekera, and E. Afgan. BioBlend: automating pipeline analyses within Galaxy and CloudMan. *Bioinformatics*, 29(13):1685–1686, Jul 2013.
26. E. P. Nawrocki, S. W. Burge, A. Bateman, J. Daub, R. Y. Eberhardt, S. R. Eddy, E. W. Floden, P. P. Gardner, T. A. Jones, J. Tate, and R. D. Finn. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, 43(Database issue):D130–137, Jan 2015. [PubMed Central:PMC4383904] [DOI:10.1093/nar/gku1063] [PubMed:25392425].
27. H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics*, 14(2):178–192, Mar 2013. [PubMed Central:PMC3603213] [DOI:10.1093/bib/bbs017] [PubMed:22517427].
28. C. Tyner, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, C. Eisenhart, C. M. Fischer, D. Gibson, J. N. Gonzalez, L. Guruvadoo, M. Haeussler, S. Heitner, A. S. Hinrichs, D. Karolchik, B. T. Lee, C. M. Lee, P. Nejad, B. J. Raney, K. R. Rosenbloom, M. L. Speir, C. Villarreal, J. Vivian, A. S. Zweig, D. Haussler, R. M. Kuhn, and W. J. Kent. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.*, 45(D1):D626–D634, Jan 2017. [PubMed Central:PMC5210591] [DOI:10.1093/nar/gkw1134] [PubMed:27899642].